Performance of Decision Trees and Artificial Neural Networks in Binary classification Problem

Kishan Shukla

Bachelor Of Technology, Indian Institute of Technology, Kanpur, India

1. Problem Statement

A company, is planning to develop a model that is capable of differentiating earning manipulators from non-manipulators for a bank that handles out commercial loans to SMEs. Technically, this is a Binary classification problem. The data available to me has 1239 examples and each example has 8 features and a single label (0 or 1).

2. Methodology

We will be trying out Decision Tree and Artificial Neural Network models for the given problem and compare the results obtained from the models. We will be trying out three different decision tree models namely CART, C4.5 and ID3 models which are different in there splitting criteria, types of data they can handle, pruning strategy, etc.

Table 1. *Different Decision Tree models*

MODEL	Splitting Criteria	Attribute type	Missing Values	Pruning Strategy
CART	Towing criteria	Handles both categorical and Numerical values	Handles missing values	Cost complexity pruning is used
C4.5	Gain ratio	Handles both categorical and Numerical values	Handles missing values	Error based pruning is done
ID3	Information gain	Handles only categorical values	Do Not Handle missing values	No pruning is done

3. Data Analysis

Total 1239 example data is available, in which 39 belong to manipulator class and 1200 belong to non manipulator class. From the data it is clear that we are facing class imbalance problem as the proportion of data is not balanced between the two classes. We can't train our models to this data as our models would be biased then. Even if our model predicts non-manipulator class for every test data then also our accuracy would be 1200/1239 (= 0.9685), which is quite high and wrong indicator of the models performance.

Therefore we need to deal with the Class Imbalance problem first then train our models. To deal with Class Imbalance, I used an Oversampling technique called SMOTE (Synthetic Minority Oversampling Technique), what it did is basically tried to predict the probability distribution of the manipulators class samples and generated more samples from the predicted distribution. After applying SMOTE we got about 480 total manipulator class samples and 1200 non-manipulator class samples.

Since the values of the features have mean around 1 and standard deviation less than 3 (shown in the code), there is no need to normalize the data.

4. Results

As accuracy is not good indicator for evaluating the models performance (due to class imbalance) I used (ROC) AUC and F1 scores as well to compare the models. The results obtained are tabulated below:

Table 2. *Score of the models*

Model	Accuracy	(ROC) AUC score	F1 score
CART DT	92.26190476190477	89.62609970674488	0.8505747126436781
C4.5 DT	87.3015873015873	82.6001955034213	0.7500000000000001
ID3 DT	86.11111111111111	83.82048905681899	0.7852760736196319
ANN	95.03968253968253	94.07804451162514	0.9097472924187726

5. Conclusion

From the scores, we can conclude that Artificial Neural Network works better than Decision Tree for the given data. Among the Decision Tree models we can say that CART works better than C4.5 and ID3, whereas C4.5 and ID3 have similar performance for the given data.